International Journal of Advanced Research in

Mechanical and Production Engineering and Development

Volume: 1 Issue: 3 27-May-2014,ISSN_NO: 2320-7256

IMPORTANCE OF WEB USAGE MINING IN MACHINE LEARNING

V.DAVID MARTIN[1], J.SHARMILA[2]
[1&2]MANONMANIAM SUNDARANAR UNIVERSITY,TIRUNELVELI.

ABSTRACT - *The World Wide Web (Web) has been providing an important and indispensable platform for receiving information and disseminating information as well as interacting with society on the Internet. With its astronomical growth over the past decade, the Web becomes huge, diverse and dynamic. The application of data mining techniques to the web is called Web Mining. Web Mining aims to discover interesting patterns in the structure, the contents and the usage of web sites. An indispensable tool for the webmaster, it has, nevertheless, a long road ahead in which visualization plays an important role. Currently, Web mining techniques has emerged as an important research area to help Web users find the information needed. This paper is an effort in analyzing the views and methodologies stated by various authors on various processes in mining the web. Web Usage Mining is the application of data mining techniques to discover interesting usage patterns from Web data, in order to understand and better serve the needs of Web-based applications. Usage data captures the identity or origin of Web users along with their browsing behavior at a Web site. Learning is the ability to improve a machine's behaviour and inference/interpretation based on training data/experience. We have a large number of criteria for categorizing learning algorithms; in practice selection of a suitable algorithm is a very complex process. Somewhere, there is an abstract notion of a learning algorithm (a neural network, a genetic algorithm, a classifier system) which is made very specific, depending on the situation in which it should work. Therefore we will investigate the applicability of a number of learning algorithms by tuning certain aspects of the algorithm*

1, INTRODUCTION

The web mining is the use of data mining techniques to automatically discover and extract information from World Wide Web documents and services. This area of research is so huge today partly due to the interest in e-commerce. This phenomenon partly creates confusion what

constitutes Web mining and when comparing research in this area. Similar to, we suggest decomposing Web mining into these subtasks, namely

1. Resource finding: the task of retrieving intended Web documents.
2. Information selection and pre-processing: automatically selecting and pre-processing specific information from retrieved Web resources.
3. Generalization: automatically discovers general patterns at individual Web sites as well as across multiple sites.
4. Analysis: Validations and/or interpretation of the mined patterns

We should also note that humans play an important role in the information or knowledge discovery process on the web since the web is an interactive medium. This is especially important for validation and/or interpretation in step 4. So, interactive query-triggered knowledge discovery is as important as the more automatic data triggered knowledge discovery. However, we exclude the knowledge discovery done manually by humans. Thus, Web mining refers to the overall process of discovering potentially useful and previously unknown information or knowledge from the web data. It implicitly covers the standard process of knowledge discovery in databases (KDD). We could simply view web mining as an extension of KDD that is applied on the Web data. From the KDD point of view, the information and knowledge terms are interchangeable. There is a close relationship between data mining, machine learning and advanced data analysis. Web mining is often associated with IR or IE. However, web mining or information discovery on the web not the same as IR or IE.

Web usage mining tries to discover the useful information from the secondary data derived from the interactions of the users while surfing on the web. It focuses on the techniques that could predict user's behavior while the user interacts with web. *M.* Spiliopoulou abstract the potential strategic aims in each domain in to mining goal as: predication of the user's behavior within the site, comparison between expected and actual web site usages, adjustment of the web site to the interests of its users. There are no definite distinctions between the web usage mining and other

two categories. In the process if data presentation of web usage mining, the web site topology will as the information sources ,Which interacts web usage mining with the web content mining and web structure mining moreover the clustering in the process of pattern discovery is a bridge to web content and structure mining from usage mining. There are lots of works have been done in the *IR*, Database, Intelligent Agents and topology, which provides a sound function for the web content, web structure mining. Web usages mining is a relative new research area, and gains more and more attentions in recent years. I will have a detailed introduction in the next section about mining, based on some up-to-date research works.

2, APPROACH OF WEB USAGE MINING

The web usage mining generally includes the following several steps: data collection, data pretreatment, knowledge discovery and pattern analysis.

A] Data collection:

Data collection is the first step of web usage mining, the data authenticity and integrality will directly affect the following works smoothly carrying on and the final recommendation of characteristic service's quality. Therefore it must use scientific, reasonable and advanced technology to gather various data. At present, towards web usage mining technology, the main data origin has three kinds: server data, client data and middle data (agent server data and package detecting).

B] Data preprocessing:

Some databases are insufficient, inconsistent and including noise. The data pretreatment is to carry on a unification transformation to those databases. The result is that the database will to become integrate and consistent, thus establish the database which may mine. In the data pretreatment work, mainly include data cleaning, user identification, session identification and path completion.

Volume: 1 Issue: 3 27-May-2014,ISSN_NO: 2320-7256

1) Data Cleaning:

The purpose of data cleaning is to eliminate irrelevant items, and these kinds of techniques are of importance for any type of web log analysis not only data mining. According to the purposes of different mining applications, irrelevant records in web access log will be eliminated during data cleaning. Since the target of Web Usage Mining is to get the user's travel patterns, following two kinds of records are unnecessary and should be removed:

1. The records of graphics, videos and the format information The records have filename suffixes of GIF,JPEG, CSS, and so on, which can found in the URI field of the every record;
2. The records with the failed HTTP status code. By examining the Status field of every record in the web access log, the records with status codes over 299 or under 200 are removed.

It should be pointed out that different from most other researches, records having value of POST or HEAD in the Method field are reserved in present study for acquiring more accurate referrer information.

Volume: 1 Issue: 3 27-May-2014,ISSN_NO: 2320-7256

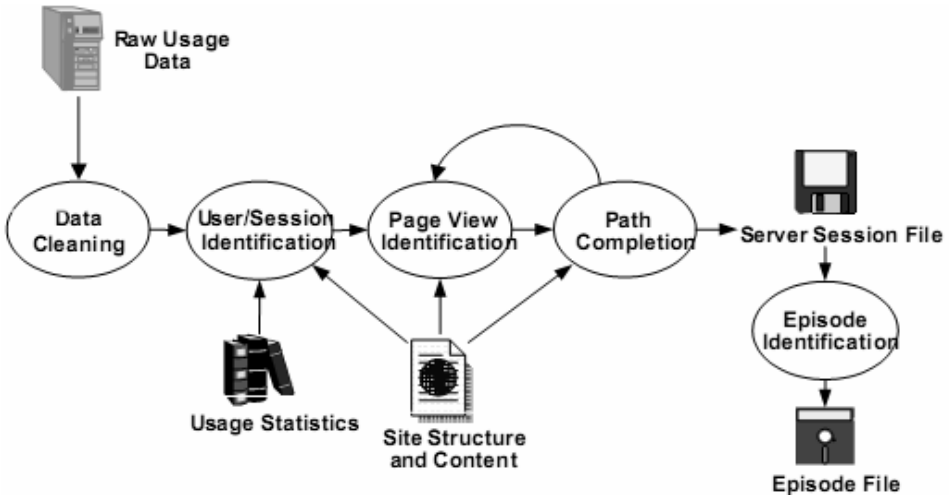

Fig. 1 Approaches of Web Usage Mining

2) User and Session Identification:

The task of user and session identification is find out the different user sessions from the original web access log. User's identification is, to identify who access web site and which pages are accessed. The goal of session identification is to divide the page accesses of each user at a time into individual sessions. A session is a series of web pages user browse in a single access. The difficulties to accomplish this step are introduced by using proxy servers, e.g. different users may have same IP address in the log. A referrer-based method is proposed to solve these problems in this study. The rules adopted to distinguish user sessions can be described as follows:

a. The different IP addresses distinguish different users;

b. If the IP addresses are same, the different browsers and operation systems indicate different users;

c. If all of the IP address, browsers and operating systems are same, the referrer information should be taken into account. The Refer URI field is checked, and a new user session is identified if the URL in the Refer URI field hasn't been accessed previously, or there is a large interval (usually more than 10 seconds) between the accessing time of this record and the previous one if the Refer URI field is empty;

d. The session identified by rule 3 may contains more than one visit by the same user at different time, the time oriented heuristics is then used to divide the different visits into different user sessions. After grouping the records in web logs into user sessions, the path completion algorithm should be used for acquiring the complete user access path.

3) Path completion

Another critical step in data preprocessing is path completion. There are some reasons that result in path's incompletion, for instance, local cache, agent cache, "post" technique and browser's "back" button can result in some important accesses not recorded in the access log file, and the number of Uniform Resource Locators(URL) recorded in log may be less than the real one. Using the local caching and proxy servers also produces the difficulties for path completion because users can access the pages in the local caching or the proxy servers caching without leaving any record in server's access log. As a result, the user access paths are incompletely preserved in the web access log. To discover user's travel pattern, the missing pages in the user access path should be appended. The purpose of the path completion is to accomplish this task. The better results of data pre-processing, we will improve the mined patterns' quality and save algorithm's running time. It is especially important to web log files, in respect that the structure of web log files are not the same as the data in database or data warehouse. They are not structured and complete due to various causations. So it is especially necessary to pre-process web log files in web usage mining. Through data pre-processing, web log can be transformed into another data structure, which is easy to be mined.

C] Knowledge Discovery

Use statistical method to carry on the analysis and mine the pretreated data. We may discover the user or the user community's interests then construct interest model. At present the usually used machine learning methods mainly have clustering, classifying, the relation discovery and the order model discovery. Each method has its own excellence and shortcomings, but the quite effective method mainly is classifying and clustering at the present.

D] Pattern analysis

Challenges of Pattern Analysis is to filter uninteresting information and to visualize and interpret the interesting patterns to the user. First delete the less significance rules or models from the interested model storehouse; Next use technology of OLAP and so on to carry on the comprehensive mining and analysis; Once more, let discovered data or knowledge be visible; Finally, provide the characteristic service to the electronic commerce website.

3, ENHANCING WEB-BASED LEARNING ENVIRONMENTS

WebSIFT is a set of comprehensive web usage tools that is able to perform many data mining tasks and discover a variety of patterns from web logs. A versatile system, WebLogMiner [16], uses data warehousing technology for pattern discovery and trend summarization from web logs. However these wide-ranging tools are not integrated in e-learning systems and it is cumbersome for an educator who doesn't have extensive knowledge in data mining to use these tools to improve the effectiveness of web-based learning environments. A new web usage mining system dedicated for e-learning is being developed to allow educators to assess on-line learning activities. For an educator using a web-based course delivery environment, it could be beneficial to track the activities happening in the course web site and extract patterns and behaviours prompting needs to change, improve, or adapt the course contents. For example, one could identify the paths frequently and regularly visited, the paths never visited, the clusters of learners based on the paths they follow, etc. For a learner using a web-based course delivery environment, it could be beneficial to receive hints from the system on what subsequent activity to perform based on similar behavior by other "successful" learners. For example, the system

Volume: 1 Issue: 3 27-May-2014,ISSN_NO: 2320-7256

could suggest shortcuts to frequently visited pages based on previous user activities, or suggest activities that made similar learners more ”successful”. It could also be beneficial if the system adapts the course content logical structure to the learner’s learning pace, interest, or previous behaviour. Web-based course content is not always presented and structured in an intuitive way. By analyzing common traversal paths of the course content web pages or frequent changes in individual traversal paths, the layout of the course can be reorganized or adapted to better fit the needs of a group or an individual. We see two types of data mining in the context of e-learning: off-line web usage mining and integrated web usage mining. Off-line web usage mining is the discovery of patterns with a standalone application. This pattern discovery process would allow educators to assess the access behaviours, validate the learning models used, evaluate the learning activities, compare learners and their access patterns, etc. We have designed and implemented a prototype of such an application as a tool for educators to apply association rules to discover relationships between learning activities that learners perform, sequential analysis to discover interesting patterns in the sequences of on-line activities, and clustering to group similar access behaviours. While most data mining algorithms need specific parameters and threshold values to tune the discovery process, the users of web usage mining applications in the context of e-learning, namely educators and e-learning site designers, are not necessarily savvy in the intricate complexities of data mining algorithms. For this purpose we have tried to design new algorithms that need minimum input from the user and automatically adjust to the web log data at hand. Here we propose a totally non-parametric approach for clustering web sessions. Off-line web usage mining helps educators put in question and validate the learning models they use as well as the structure of the web site as it is perused by the learners. In contrast, integrated web usage mining is a process of discovering patterns that is incorporated with the e-learning application. This encompasses adaptive web sites, personalization of activities, and automatic recommenders that suggest activities to learners based on their preferences as well as their history of activities and the access patterns discovered from the communal accesses. We are currently designing a recommender based association rule mining similar to the text categorization we developed. The idea consists of discovering relevant associations between learning activities and generating association rules that are applied in real time when in a current

Volume: 1 Issue: 3 27-May-2014,ISSN_NO: 2320-7256

session the activities of the antecedent of a rule are verified then the activities in the consequent of the rule are suggested to the learner as the recommended next step in the learning session. The algorithm for text categorization presented can also be used to automatically categorize learners' messages sent on an asynchronous conferencing system in order to help the educators better assess the information exchange in a course related forum.

CONCLUSIONS AND FUTURE WORK

The Web is an excellent tool to deliver on-line courses in the context of distance education. However, counting only on web traffic statistical analysis does not take advantage in the potential of hidden patterns inside the web logs. Web usage mining is a non-trivial process of extracting useful implicit and previously unknown patterns from the usage of the Web. Significant research is invested to discover these useful patterns to increase profitability of e-commerce sites. However, the goals of these applications and methods, "turning visitors into purchasers", are different from the goals in e-learning: "turning learners into effective better learners." We have seen some examples where data mining techniques can enhance on-line education for the educators as well as the learners. While some tools using data mining techniques to help educators and learners are being developed, the research is still in its infancy. In addition, with the awareness of the potential advantages of integrated web usage mining and the insufficient data recorded by web servers, there is a need for more specialized logs from the application side to enrich the information already logged by the web server. This added value by specific event recording on the e-learning side will give clicksteams and the patterns discovered a better meaning and interpretation.

REFERENCES

[1] Tak-Lam Wong and Wai Lam, “Learning to Adapt Web Information Extraction Knowledge and Discovering New Attributes via a Bayesian Approach”, IEEE Transactions On Knowledge And Data Engineering, vol. 22, no. 4, pp: 523-536, 2010.

[2] Yatsko V., Shilov S. and Vishniakov T., “A Semi-automatic Text Summarization System”, In proceedings of the 10th International Conference on Speech and Computer, Patras, pp. 283-288, 2005.

[3] LaddaSuanmali, NaomieSalim and Mohammed Salem Binwahlan, "Automatic Text Summarization Using Feature Based Fuzzy Extraction", vol. 20, no. 2, pp. 105-115, November 2009.

[4] KaustubhPatil and PavelBrazdil, "SUMGRAPH: Text Summarization Using Centrality In The Pathfinder Network", International Journal on Computer Science and Information Systems, vol.2, no.1, pp. 18-32, 2007.

[5] RachitArora and BalaramanRavindran, “Latent Dirichlet Allocation Based Multi-Document Summarization”, In Proceedings of the second workshop on Analytics for noisy unstructured text data, pp:91-97, 2008.

[6] KhosrowKaikhah, “Automatic Text Summarization with NeuralNetworks”, Second IEEE International conference on intelligent systems, pp: 40-45, 2004.

[7] H. Edmundson,“New methods in automatic extracting”, Journal of the Association for Computing Machinery, Vol: 16, No. 2, pp: 264-285, 1969.

[8] Inderjeet Mani, “Recent Developments in Text Summarization", In Proceedings of the tenth international conference on Information and knowledge management, ACM Press, pp: 529 - 531, 2001

[9] ShiyanOu, Christopher S.G. Khoo and Dion H. Goh, "Design and development of a concept-based multidocument summarization system for research abstracts", Journal of Information Science, vol. 34 , no. 3, pp. 308-326 , June 2008.